Multi-Object Recognition Method Inspired by Multimodal Information Processing in the Human Brain

Ryoga Seki
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
r-seki@ist.osaka-u.ac.jp

Daichi Kominami
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
d-kominami@ist.osaka-u.ac.jp

Hideyuki Shimonishi
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
and NEC Corp.
h-shimonishi@ist.osaka-u.ac.jp

Masayuki Murata
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
murata@ist.osaka-u.ac.jp

Masaya Fujiwaka
Data Science Research Labs, NEC Corporation
Kanagawa, Japan
fujiwaka@nec.com

***Abstract*—In order to realize the Digital Twin, it is necessary to instantly understand various objects existing in the real world through various sensor devices. In recent years, the development of machine learning technologies such as a convolutional neural network has been remarkable, and in the field of video analysis, they achieve very high recognition rates. However, if all sensor data were collected and processed in the cloud, network transmission bandwidth and communication delays would become bottlenecks. The brain is a light weight system that makes decisions based on uncertain observations and We have previously proposed an object recognition method based on a mathematical model of how the brain recognizes information. In this paper, we extend this method. The features of the brain's recognition of spatial context are modeled by a conditional random field and incorporated into it. We show that our proposal can recognize multiple objects with an accuracy of more than 83.2% even from noisy information, and can be applied to 60 fps video in the evaluation environment.**

***Index Terms*—Mobile AR, digital twin, multimodal recognition, Bayesian attractor model, Bayesian causal inference.**

I. Introduction

Internet of Things (IoT) [1] has been attracting attention that links various devices with network services. Through the IoT network, the positions and states of objects in the real world can be measured and stored on a computer, which allows for more diverse services. In this way, all objects in the real world can be identified and mapped onto a virtual space, which is called a *Digital Twin* [2]. When the digital twin is realized, every event in the real world can be simulated in the virtual world, and the feedback is expected to be useful in various situations, such as ensuring safety and saving resources [3].

In order to construct a wide-area real-world digital twin in real time, video from many cameras must be processed in real time. If all video data were collected and processed in the cloud, network transmission bandwidth and communication delay would become bottlenecks. Therefore, it is desirable to construct a digital twin in which video data is analyzed at the edge device in the immediate vicinity of the camera, and only the object data extracted from the analysis is collected in the cloud. However, current high-precision CNNs require a large amount of computing power [4], making real-time analysis at the edge difficult, and recognition accuracy is sacrificed when computational power is reduced by a lightweight model.

A familiar example of a lightweight system that makes decisions from uncertain observational information is the information processing mechanism of the brain. Moreover, it is known that the human brain is not only light-weight but also processes cognition through multimodal information processing [5]–[7], and there are efforts to improve the accuracy of cognition by machine learning with reference to such a mechanism [8], [9]. The brain uses uncertain information obtained from the eyes, ears, skin, and semicircular canals to infer the state of the surrounding environment and to make final decisions. Such a brain recognition mechanism is a lightweight system that can conserve computational resources. Information processing of the brain was modeled hierarchically in [10]; feature, unimodal, and multimodal levels. In [11], we previously proposed an object recognition method that employs this hierarchical model and for the unimodal and multimodal levels, we use the Bayesian attractor model (BAM) [12] and Bayesian causal inference (BCI) [5]–[7], [13] respectively, while basic media-specific data preprocessing is performed at the feature level.

The BAM represents the behavior of a person's decision-making process based on observed information by using Bayesian estimation, and thus, it is expected to identify objects with high accuracy from uncertain time-varying information. BCI is a mathematical model of the process by which humans recognize perceptual objects using multiple modalities (e.g., vision and audition). In this cognitive model, humans infer whether two input stimuli originate from the same stimulus source, and then integrate each input stimulus to make a final cognitive decision.

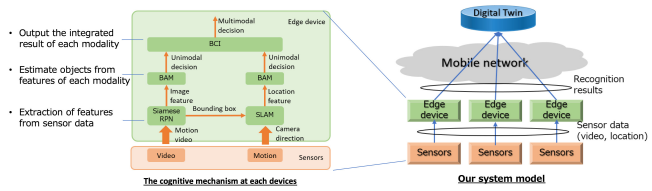

Fig. 1. Object recognition inspired by the human brain

Figure 1 shows our system model. Small edge devices equipped with multiple sensors, such as video cameras, depth cameras, LiDAR, motion sensors, etc. are distributed over a network. As described earlier, object recognition tasks are not performed in a cloud but distributed to edge devices, which read sensor data and recognition tasks described below, and send recognition results to the cloud. In the cloud, recognition results from various devices are integrated and maintained as a digital twin. At each edge device, by using the BAM to estimate objects individually from the features of the video modality and the location modality obtained from the sensor devices, and then integrating them with BCI to make decisions.

In this paper, we extend our previous method by incorporating features that recognize spatial and temporal context. Our previous method outputs a label indicating what the object is based on the previously stored image of the object and the object's location in the object recognition task. The higher-level recognition performed by the brain allows it to grasp spatio-temporal information about an object. For example, there are no multiple identical objects in space, and the distance they move in time within one frame of video is very small. For this recognition, we used a conditional random field (CRF) that is capable of capturing spatio-temporal information [14].

The remainder of this paper is organized as follows. In Section II, we describe the mathematical basis of the BAM and BCI. In Section III, we describe a method for applying the results of extracting the features of each modality from video and inputting them into the BAM and then the BCI model. In Section IV, we evaluate our proposed method. Section V gives the conclusion of this paper.

II. Related Work

A. Bayesian Attractor Model

The BAM estimates which of the pre-stored options matches the observation target. It has been applied to the research of networks in [15], [16]. The BAM has an internal decision state $\mathbf{z}_t$, and it updates this state when receiving external observations $\mathbf{x}_t$. By updating the state based on Bayesian inference, $\mathbf{z}_t$ is not treated as a single point, but is represented as a probability distribution $P(\mathbf{z}_t)$ that reflects the uncertainty of the observation and brain state.

We prepare as many stable points (attractors) $\phi_1, \ldots, \phi_n$ in the state space where $\mathbf{z}$ exists as the number of choices (n), and make the decision to take the i-th choice when $\mathbf{z}$ is sufficiently close to ϕ. Because $\mathbf{z}_t$ is represented as a probability distribution, we derive a probability density (hereafter called the *confidence* level) where $\mathbf{z}_t = \phi_i$ and make a decision using this confidence level. The details of state updating and decision making are described below.

1) State Update: The state update is performed by finding the posterior distribution $P(\mathbf{z}_t|\mathbf{x}_t)$ of the decision state $\mathbf{z}_t$ by Bayesian estimation when the observed value $\mathbf{x}_t$ is obtained. The following generative model is assumed for $\mathbf{x}_t$ and $\mathbf{z}_t$:

$$\mathbf{z}_t - \mathbf{z}_{t-\Delta t} = \Delta t f(\mathbf{z}_{t-\Delta t}) + \sqrt{\Delta t}\mathbf{w}_t, \quad (1)$$

$$\mathbf{x}_t = M\sigma(\mathbf{z}_t) + \mathbf{v}_t, \quad (2)$$

where $f(\mathbf{z})$ represents the dynamics of a Hopfield network, one of the attractor models, and the dynamics has multiple attractors. When n is the number of options to be stored in the Bayesian attractor model, we design it so that f has n attractors $\phi_1, \ldots, \phi_n$. In the above equations, M is a feature matrix for each option, with $M = [\mu_1, \ldots, \mu_n]$ where μ_i is a feature vector of the i-th option; σ is a multidimensional sigmoid function with a value range of 0 to 1; and $\mathbf{w}_t$ and $\mathbf{v}_t$ are random numbers following normal distributions $\mathbf{w}_t \sim \mathcal{N}(0, \frac{q^2}{\Delta t} I)$ and $\mathbf{v}_t \sim \mathcal{N}(0, r^2 I)$, respectively. $\mathcal{N}$ denotes a normal distribution. I is a unit matrix, and when Δt is 1, the respective standard deviations are q and r. These deviations determine the noise level of the dynamics and observation in the generative model, so q is called the dynamics uncertainty and r is called the sensory uncertainty.

2) Decision Making: Which attractor the internal state of the brain is close to can be determined based on the magnitude of the confidence level and therefore, decision making is done when the confidence level for one of the options exceeds the threshold value. The posterior probability distribution of $\mathbf{z}_t$, $P(\mathbf{z}_t|\mathbf{x}_t)$, is obtained by the state update. Here, approximate calculations are performed using the unscented Kalman filter (UKF) to account for the non-linearity of the generative model, as given in [12]. Finally, the confidence level is obtained for each pre-stored option, $\mathbf{c} = \{c_1, \ldots, c_n\}$, and $c_i = P(\mathbf{z}_t = \phi_i|\mathbf{x}_t)$. Note that since the Kalman filter is used, estimated $P(\mathbf{z}_t|\mathbf{x}_t)$ is probability density function of a multivariate normal distribution.

B. Bayesian Causal Inference

BCI is a mathematical model of human cognition based on multimodal perceptual stimuli. For example, when we see something on the left side of our field of vision and hear a sound from the same direction, we may judge the direction of arrival by integrating our visual and auditory senses, assuming that these stimuli originate from the same source, or we may judge the direction of arrival separately, assuming that they come from different directions. A comprehensive knowledge of causal inference is summarized in [17], and its application to information science fields such as machine learning is discussed in [18].

In [5], a mathematical formulation for the following process is given. When an object is presented, both modalities (visual and auditory) perceive the location of the object and probabilistically infer whether both modalities observe the same object (causal inference), which is used to integrate

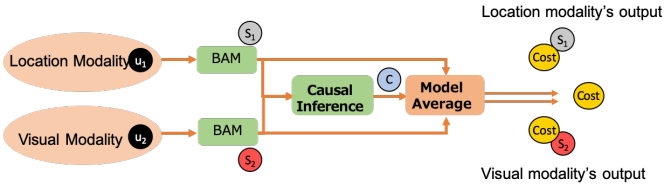

Fig. 2. Extending the BAM with BCI

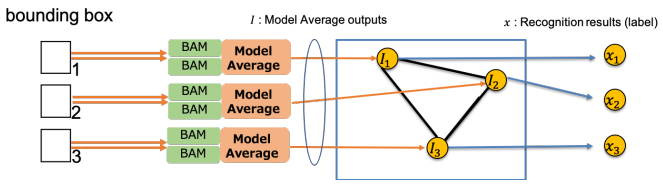

Fig. 3. Example of CRF graph structure

the observations from each modality to recognize the object's location (*model average*). The result is a weighted sum of the integrated recognition result and segregated recognition results. If the causal inference obtains a higher confidence that both modalities observe the same object, model-average results place a higher weight on the integrated result; otherwise, model-average results place a higher weight on the segregated results.

C. Object Recognition method with BAM and BCI

Here, we describe our previous work [11]. As shown in Figure 2, we integrated the BAM and the BCI, where the confidence level output from the BAM, **c**, is used as an input to the BCI, causal inference is performed to infer whether the same object is observed in the video and location modalities, and the result is used for multimodal integration by the model average algorithm to output the final object.

To use the BAM and the BCI for object estimation, we stored the features about the object to be recognized, taken from the two modalities, the video modality and the location modality, in the BAM by assigning them to a matrix M. The features used in each modality are described below.

Video-modality features: The features are extracted using a Siamese region proposal network (RPN) [19]. The Siamese RPN takes a template image and a detection image as input, detects similarities between these images, and outputs the location as a bounding box. The Siamese RPN consists of a Siamese neural network and an RPN. The former extracts the features of the image, and the latter uses the extracted features to calculate the similarities to the template image from the detected image. In the literature [19], existing CNNs such as AlexNet [4] are used as the Siamese network, but in the proposed method, a simple CNN consisting of four layers is used to greatly reduce the computational cost. Because the recognition decision is made by the BAM, the role of the CNN is just to extract feature values, and an encoder like a shallow CNN is suitable for this purpose. The feature values are 128-dimensional data from the output of the Siamese network (image features) corresponding to the bounding box output by the RPN.

Location-modality features: The feature values of the location modality are the 3D world coordinate system data calculated from the camera direction vector and the depth information integrated from multiple frames. Thus, the feature values to be fed to the BAM are 3D data, such as the x-y-z location in a world coordinate system.

Note that for both modality, to enable one-shot learning, that is, to use just one representative image and eliminate pre-training, only the first image of an object in the video where each object is first seen is used to calculate the feature values.

III. Method

The BAM makes a decision for each frame of the video and location observation in chronological order, and outputs the confidence level for each attractor. The BAM makes a decision for each frame in chronological order for each video and location observation, and outputs the confidence level for each attractor. In *unimodal object recognition*, the object with the highest confidence is the recognition result. The confidence level is given to the input of the BCI, and the decision is made by performing Causal Inference. In *multimodal object recognition*, in the model average process, the object with the lowest cost is output as the recognition result [11]. In *object recognition with CRF*, BCI outputs are used as inputs of a CRF that outputs the recognition result.

Then we will discuss how to recognize multiple objects considering spatial contexts with a CRF. From a single bounding box, a number of features are extracted for each modality, which are processed by the BAM and model averaged by the BCI. In other words, we get one output as a vector from one bounding box. To make object recognition, we build a conditional random field (CRF) with outputs of the BCI as nodes, and it decides recognition results as a label corresponding to each bounding box.

A method for assigning objects in a video to nodes in a CRF graph is described in [20]. We also define an energy function on the labels of objects that can be minimized to obtain the optimal solution. We use the Belief Propagation to find an approximate solution in this paper.

To begin with, we set the following constraints. They are not appropriate for real-world environments and need to be changed in our future studies. Then, the graph we constructed is shown in the figure 3. We explain how the CRF recognizes multi-objects from BCI output below.

- The input data is streamed without any loss.
- The input data has a one-to-one correspondence with the attractor.
- Each attractor stores a different object.

1) Energy Function: This CRF explores x that minimizes the Gibbs energy $E(x)$ for the input I to output a good

recognition result x with more correct answers. The energy function is as follow:

$$E(x|I) = \sum_i \psi_u(x_i|I_i) + \sum_{i<j} \psi_p(x_i, x_j), \quad (3)$$

where I_i is BCI's estimation of the features obtained from the i-th bounding box, x_i is the label to be the i-th output of the CRF, $\psi_u(x_i|I_i)$ is the unary potential which represents the difference between the input I_i and x_i, and $\psi_p(x_i, x_j)$ is the pairwise potential which represents various constraints of objects. We define the unary and pairwise positions as follows:

$$\psi_u(x_i|I_i) = KLD(U_{x_i} \| I_i^{vis}) + KLD(U_{x_i} \| I_i^{loc}), \quad (4)$$

$$\psi_p(x_i, x_j) = (1 - |x_i - x_j|)\theta, \quad (5)$$

where $KLD(P\|Q)$ is Kullback–Leibler divergence [21] which is a measure of the difference between two probability distributions, $|x_i - x_j|$ is 0 if $x_i = x_j$, and otherwise it is 1, U_{x_i} is a vector $(0 \ldots 0, 1, 0 \ldots 0)$ where the x_i-th element is 1 and the others are 0, and θ is the penalty when $x_i = x_j$ despite the constraint that there are no objects that are the same. From the above definitions, the x that minimizes the energy function defined in Equation 3 is the appropriate solution for input I.

2) Belief Propagation: There are many ways to find the minimum value of a function, but in this paper, we have adopted the Belief Propagation method [22]. In [22], it is shown that the Belief Propagation method was used to find the optimal solution for CRF. In this method, each node sends a message to its neighbors, and updates its own state. In detail, the algorithm can be divided into the following steps.

1) Determine the initial value of the node i's state x_i and messages $m_{i,i}(x_i)$ to itself by calculating only the unary potential only.
2) A message $\tilde{m}_{i,j}(x_j)$ is sent from node i to node j.
3) Update the state of node i to x_i^*.

As the constraints become more complex, we need to repeat steps 2. and 3. multiple times. In step 2, the calculation of the message $\tilde{m}_{i,j}(x_j)$ is done as follow:

$$\tilde{m}_{i,j}(x_j) = \min_{x_i}(\psi_u(x_i|I_i) + \psi_p(x_i, x_j) + \sum_{k \in N_i \setminus j} \tilde{m}_{k,i}(x_i)), \quad (6)$$

and we defined the rule updating x_i^* as follow:

$$x_i^* = \arg\min_{x_i}(\psi_u(x_i|I_i) + \sum_{k \in N_i} \tilde{m}_{k,i}(x_i)). \quad (7)$$

IV. Evaluation

Here, we assume the system shown in Figure 1, and evaluate the proposed method as one that performs video analysis processing at the edge.

A. Environment

As for the computational time of the proposed method, the most computationally intensive part is the part where 128-dimensional video features are input to the BAM, $\mathbf{z}_t$ is estimated, and the confidence level is output. We use a desktop PC (CPU: Core-i7 10750H, RAM: 16.0 GB) for feature extraction of data obtained from sensor instruments and a desktop PC (CPU: Core-i7 8700, RAM: 16.0 GB) for recognition in BAM.

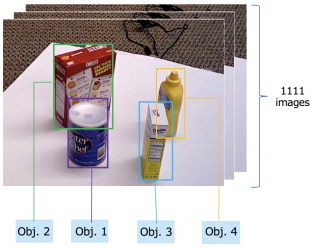

Fig. 4. 4 Objects video data

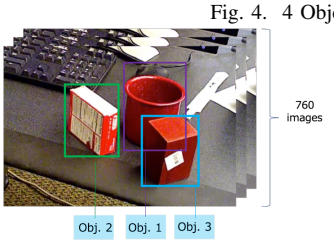

Fig. 5. 3 Objects video data

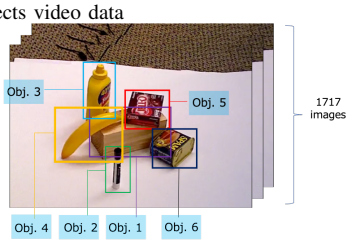

Fig. 6. 6 Objects video data

In our envisioned digital twin concept, local virtual worlds are connected to form a global digital twin. The local world here refers to the status of a small space, like a single room, where local real-time interactions are expected, thus data such as the position and status of local objects are maintained locally. In this research environment, feature extraction and recognition are performed by different devices. In this research environment, feature extraction and recognition are performed by different devices. Feature extraction is performed at each sensor terminal, and recognition is performed at the edge device in each room.

For the video dataset, we used a real measurement public dataset of various objects (Yale-CMU-Berkeley Object and Model set) [23]. Figure 4 shows the video data used for input. For each frame and each of four objects in the video data, we extracted the features of the video modality and location modality using the method in Section II-C a). As for the two parameters of the BAM, (q, r), we set (0.3, 2.5) for the video modality and (1.5, 0.04) for the location modality. In the following evaluation, for each object, we determined whether the object was correctly identified when the video modality and location modality features were observed. For this purpose, the BAM stored the feature values of each of the video and location modalities for the four objects. Similarly, Figs. 5 and 6 show video data with three and six objects in each image, respectively. We used these three video data sets to evaluate the proposed method.

For the video modality in this evaluation, we assumed one-shot learning for AR applications, and thus only the first image of the object in the video was used as training data, and thus, we cannot compare directly our results and existing CNN

based object classification methods. We can compare our BAM output of the video modality and the output from classification branch of Siamese RPN. In the latter case, recognition results are given for each video frame independently without considering other frames in the sequence, so this would suffer from recognizing an object from motion video data than our method using BAM. Therefore, in this evaluation, we compared our method with Oracle data, i.e. expected results.

B. Result

1) Unimodal Object Recognition: Figures 7 shows the confidence outputs of the BAM. In these figures, the left Confidence Modal 1 represents the video modality, and the right Confidence Modal 2 represents the location modality. The same is true for the modalities in the subsequent figures. The object with the highest confidence is used as the identification result. For 4-objects recognition, the correct response rate of the BAM with video modality only was 79.41%, and that with location modality only was 81.66%, where the correct response rate is defined as the percentage of identification results output per input that match the observed object. The 4-objects recognition, are summarized in the unimodal row in Tables I, II and III. In Fig. 7, the vertical axis is the confidence level in normal logarithm, which is the result of a decision about which object is currently being observed, and the horizontal axis is the time step. The output of the 3-object recognition is stable, but we can see that it becomes unstable as the number of objects increases. Next, we will focus on whether this unstable output can be improved by multimodal recognition.

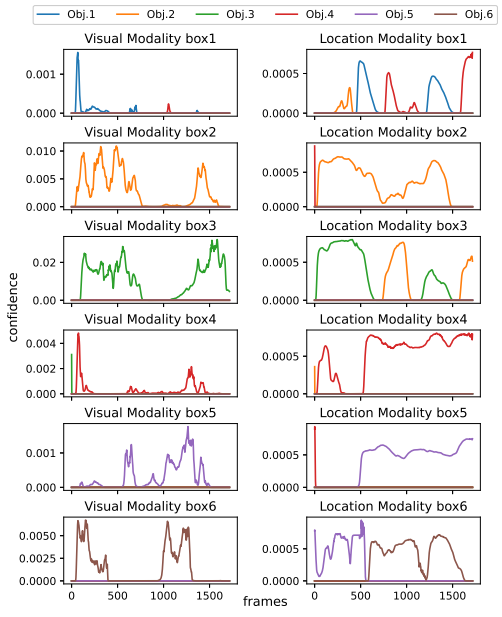

Fig. 7. Confidence of 6 Objects

2) Multimodal Object Recognition: The confidence levels in the video and location modalities, obtained in Section IV-B1, are given to the BCI to evaluate the results of integrating Causal Inference and discrimination results. In Tables I, II and III, each "Multimodal" modality shows the percentage of correct responses resulting from performing Model Average on each modality. In Model Average, the output is based on one modality while the other modality complements the recognition result. We do not discuss in this paper which modality should be treated as the main one in multimodal, and consider it as a future issue.

Obj. 1 in Table I and Obj. 1 in Table II have a high rate of correct answers in unimodal recognition and show little change, but when only one of the modalities has a high rate of correct answers, there are improvements in the recognition results. It is also important that the recognition of modalities that were originally good has not been affected much for the worse.

3) Object Recognition with CRF: To the above multimodal recognition result, we further impose the constraint "no two objects are the same" to make it multi-object recognition. These final percentages of correct answers are summarized in Tables. I, II, and III. For three- and four-object recognition, sufficient accuracy has already been obtained at the unimodal and multimodal levels. This is because there is little overlap between objects in a video frame and the observed values are stable. In such cases, imposing too strong constraints and their penalties will adversely affect the recognition accuracy. For example, if the correct answer label is $(1,\ 2,\ 3,\ 4)$, and the BCI answers $(1,\ 1,\ 3,\ 4)$ without successfully recognizing the second object, the recognition result may be set to $(2,\ 1,\ 3,\ 4)$ due to a penalty. This has not resulted in a significant decrease in the rate of correct answers, but it should be kept in mind.

On the other hand, the recognition result of 6 objects is positive. As can be seen from Fig. 6, there is a lot of overlap between objects, and there are objects for which the observed values are not accurate to begin with. It can be read from Table. III that such objects are helped by multimodal recognition to recognize other objects. We can see that objects that could not be recognized properly even with multi modality are now recognized correctly.Although the recognition accuracy is a little lower than the unimodal, it is less accurate than the average of the two modalities.

C. System level evaluation

We have confirmed that the combination of multiple modalities can make up for the inaccuracy of the unimodal method and achieve more accurate decisions. The actual computation time is 1.18 ms per frame input, applicable to 30 fps and 60 fps video. While the video features extracted from the sensor data had 128 dimensions, the confidence level output to the cloud is the number of attractors (6 at most in this environment). We have shown that we can significantly reduce the amount of data sent to the cloud.

V. Conclusion

We have created a flexible system that can compensate for the instability of one source of information with data from another source. We incorporated features of the brain's

TABLE I
3 OBJECTS CORRECT RESPONSE RATE (%)

Modality	Obj. 1	Obj. 2	Obj. 3	Total
Unimodal video	99.7	85.0	68.3	84.3
Unimodal location	99.6	98.5	39.2	79.1
Multimodal video	99.6	93.4	94.7	95.9
Multimodal location	99.6	97.8	39.2	78.9
CRF version	97.4	96.8	96.8	97.0

TABLE II
4 OBJECTS CORRECT RESPONSE RATE (%)

Modality	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Total
Unimodal video	100.0%	25.2%	97.8%	94.6%	79.4%
Unimodal location	99.6%	97.7%	29.6%	99.7%	81.7%
Multimodal video	99.6%	82.1%	98.5%	98.5%	93.8%
Multimodal location	99.6%	97.7%	36.6%	99.7%	83.4%
CRF version	97.7%	97.7%	98.3%	98.3%	98.0%

TABLE III
6 OBJECTS CORRECT RESPONSE RATE (%)

Modality	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Total
Unimodal video	85.3	99.4	95.9	91.7	96.9	73.7	90.5
Unimodal location	39.8	92.6	67.8	92.0	74.6	66.6	72.2
Multimodal video	75.2	99.4	96.5	94.8	96.9	83.0	91.0
Multimodal location	39.8	92.7	67.9	92.2	74.8	66.6	72.3
CRF version	94.5	83.8	83.2	92.4	86.6	86.8	87.9

recognition of spatial and temporal context by conditional probability fields.First, we discuss processing speed. As the number of objects to be recognized increases, the number of bounding boxes increases, and each BAM and BCI will run in parallel to create a larger graph. In this paper, the nodes in the graph are connected to all the others, which increases the exchange of messages. We need to work out how to create graphs and how to connect nodes. Specifically, it could be to create pairs of nodes that do not need to be connected with new constraints. The way constraints are defined may also need to be automated. And we consider the percentage of correct answers. The recognition accuracy of BAM is significantly degraded if incorrect data is stored in the attractor.The same is true if there are too many attractors. In this paper, the first video frame was memorized, but we need to improve this. There are several possible ways to improve it, such as getting an average of several frames, or updating the attractor over time. On another note, we need to build a higher level of cognitive layer in order to realize the digital twin. Figure 1 shows the structure of the digital twin as we envision it. The configuration of the edge devices has been discussed in this paper, but there is no idea yet of a higher-level system that integrates the information obtained from each edge. Simulation with a built-in temporal context is also a future task.

ACKNOWLEDGEMENT

A part of this work was supported by National Institute of Information and Communications Technology (NICT) in Japan.

REFERENCES

[1] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Business horizons*, vol. 58, no. 4, pp. 431–440, 2015.

[2] F. Tao, H. Zhang, A. Liu, and A. Y. Nee, "Digital twin in industry: State-of-the-art," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2018.

[3] C. Boje, A. Guerriero, S. Kubicki, and Y. Rezgui, "Towards a semantic construction digital twin: Directions for future research," *Automation in Construction*, vol. 114, 2020.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[5] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisensory perception," *PLoS one*, vol. 2, no. 9, p. e943, 2007.

[6] M. Aller and U. Noppeney, "To integrate or not to integrate: Temporal dynamics of hierarchical bayesian causal inference," *PLoS biology*, vol. 17, no. 4, p. e3000210, 2019.

[7] R. L. French and G. C. DeAngelis, "Multisensory neural processing: from cue integration to causal inference," *Current Opinion in Physiology*, pp. 8–13, Aug. 2020.

[8] S. T. Namin, M. Najafi, M. Salzmann, and L. Petersson, "Cutting edge: Soft correspondences in multimodal scene parsing," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[9] M. Subedar, R. Krishnan, P. L. Meyer, O. Tickoo, and J. Huang, "Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] R. Velik, "A brain-inspired multimodal data mining approach for human activity recognition in elderly homes," *Journal of Ambient Intelligence and Smart Environments*, vol. 6, no. 4, pp. 447–468, 2014.

[11] R. Seki, D. Kominami, H. Shimonishi, M. Murata, and M. Fujiwaka, "Object estimation method for edge devices inspired by multimodal information processing in the brain," in *19th IEEE Annual Consumer Communications & Networking Conference (CCNC) (poster session)*. IEEE, January 2022.

[12] S. Bitzer, J. Bruineberg, and S. J. Kiebel, "A Bayesian attractor model for perceptual decision making," *PLoS Computational Biology*, vol. 11, no. 8, p. e1004442, 2015.

[13] T. Rohe and U. Noppeney, "Cortical hierarchies perform bayesian causal inference in multisensory perception," *PLoS Biol*, vol. 13, no. 2, p. e1002073, 2015.

[14] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 109–117, December 2011.

[15] O. Alparslan and S. Arakawa, "Fast/slow-pathway bayesian attractor model for iot networks based on software-defined networking with virtual network slicing," in *Fluctuation-Induced Network Control and Learning*. Springer, 2021, pp. 135–154.

[16] T. Otoshi, S. Arakawa, M. Murata, K. Wang, T. Hosomi, and T. Kanoh, "Flexible updating of attractors in virtual network topology control with bayesian attractor model," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[17] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[18] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[19] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[20] S.-R. Shiang, A. Gershman, and J. H. Oh, "A generalized model for multimodal perception," in *Proceedings of AAAI '17 Fall Symposium*, November 2017.

[21] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[22] L. Landrieu, C. Mallet, and M. Weinmann, "Comparison of belief propagation and graph-cut approaches for contextual classification of 3d lidar point cloud data," in *Proceedings of 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 2768–2771.

[23] "YCB benchmarks–object and model set," available at http://www.ycbbenchmarks.com/.